

LOOKING FOR CENTRAL TENDENCIES IN THE CONFORMATIONAL FREEDOM OF PROTEINS USING NMR MEASUREMENTS

FABRIZIO CLARELLI⁽¹⁾ AND LUCA SGHERI^(1,*)

ABSTRACT. We study the conformational freedom of a protein made by two rigid domains connected by a flexible linker. The conformational freedom is represented as an unknown probability distribution on the space of allowed states. A new algorithm for the calculation of the Maximum Allowable Probability is proposed, which can be extended to any type of measurements. In this paper we use Pseudo Contact Shifts and Residual Dipolar Coupling. We reconstruct a single central tendency in the distribution and discuss in depth the results.

1. INTRODUCTION

Flexibility is a key point in the functioning of macromolecules such as proteins [13, 15]. One of the few techniques which permit to extract information about the conformational freedom of proteins in physiological condition is NMR spectroscopy. In the last decade a vast literature has flourished on this topic, we refer the reader to the two recent review papers [30, 29] for a general discussion on the available techniques.

Since the temporal scale of the fluctuations in the fold of a protein is several orders smaller than the time needed to take NMR measurements [25], information about the flexibility can be only be recovered as a probability function on the set of allowed states. This set can be parametrized using for instance Cartesian coordinates of atoms involving measurements, dihedral angles of the backbone or Euler transformations determining the position of rigid protein domains.

The recovery of this probability distribution is an under-determined ill-posed problem independently of the chosen parameters. The number of constraints is in fact far too small to determine a unique solution, except in trivial cases. Without further assumptions, any set of values of the parameters which is compatible with the measurements may be seen as a solution, and in principle there is no way of telling if a solution is better than another.

The lack of uniqueness, combined with the heterogeneity of the NMR measurement scenarios, led to a plethora of different techniques with different acronyms all trying to determine the "*best*" solution. The two cited review papers try to classify (each from a different point of view) these approaches.

From the mathematical point of view it may be observed that there are two limit cases for the solutions.

- A first approach is to find the solution which minimizes the additional hypotheses on the data, thus using either the Maximum Entropy Principle (MEP) [17], or the Kullback-Leibler divergence [19], which is the relative entropy. The MEP solution maximizes the uncertainty on the data, so each feature shown by the MEP solution is relevant. On the other hand the MEP solution is in general a continuous probability distribution, so that a large number of states is normally needed in order to approximate it [11]. The large number of variables involved raises not easy computational issues.
- A second approach is to find if the measurements carry some preference for certain states. The Maximal Allowable Probability (MAP) is the largest weight of a given state in a probability distribution which is a solution. As a function of the state, the MAP is not a solution but a sharp bound from above. The zones with a large MAP are the positions favoured by the data. On the other hand it is not possible to establish to what extent the true distribution shows these asymmetries. We can only say that the largest asymmetries should be in favour of the zones indicated by the MAP technique.

Both approaches are equally able to recover the unknown probability distribution in the extremal cases. When there is very little conformational freedom the physical situation can be thought as a series of oscillations around a central state. In this case the MEP solution tends to a Dirac function of the central state, while the MAP estimate tends to be 1 for the central state and 0 for the other states. On the other hand when there is a very large conformational freedom the MEP solution tends to the uniform distribution in the space metric and the spread of the MAP estimate is minimal.

In the in-between cases the two approaches diverge. The MEP solution is obtained as the solution with the minimal spread in the probability density. On the other hand the MAP estimate for each state is obtained via solutions with the largest possible spread between the probability of the estimated state and the probabilities of the states needed to complete the solution. Since the problem is underdetermined, both approaches are consistent with the data. They focus on

different aspects of the problem and they are in some sense complementary. For a deeper discussion of this topic, we refer the reader to [29].

In this paper we show a development of the MAP approach which permits the combination of different NMR constraints. The MAP approach has been inspired by [7], and the rigorous mathematical definition of the MAP has been given in [14], though different names have been subsequently used for this same bound of the probability. A geometrical algorithm has been developed in [21] to calculate the MAP estimate when only Residual Dipolar Coupling (RDC) [37] are considered, using the linearity properties of these measurements. Residual Dipolar Coupling and Pseudo Contact Shift (PCS) [18], which are frequently obtained together, have been used to analyse the conformational freedom of calmodulin in [10], using a complicated and time-consuming numerical procedure. The main difficulty is that PCS (as is the case for most sets of data) do not possess the linearity properties of RDC, which permits working on averaged tensors.

The Maximum Occurrence algorithm [9] uses a predetermined pool of conformers to calculate the maximal probability. The choice of a finite number of conformers simplifies the algorithm and reduces the time needed for the calculations. With this choice the positions which would cause physical violations of the atoms of the two domains may be directly eliminated from the sample.

The SES (Sparse Ensemble Selection) method has been developed in [4], and is focused on recovering a small set of conformers with large probabilities. A recent paper compares the two approaches and shows the information content of RDC and PCS [3].

In this paper we extend the geometrical algorithm of [21] to the case of PCS. Indeed, since we drop the linearity requirement only fulfilled by paramagnetic RDC, the approach can be extended to any set of measurements.

2. THEORY

We use the calmodulin measurement scenario [10] as our test case. Calmodulin is a protein made by two rigid domains (the N and C terminals) connected by a flexible linker, see figure 2 in section 4. A paramagnetic ion may be inserted in the binding sites of the N terminal, which is also called the metal domain. We can then measure NMR data for atoms belonging to both the N and C terminals.

The RDC measurements [37] are defined by

$$(1) \quad \delta_{rdc,j} = \frac{c_{rdc}}{\|P_j\|^5} P_j^t \chi P_j$$

where $P_j = P_{j_1} - P_{j_2}$ is the vector connecting selected pairs j_1 and j_2 of chemically linked atoms and c_{rdc} is a constant. The paramagnetic tensor χ is a symmetric and null-trace 3×3 matrix, thus depending on 5 coefficients. Since the atoms are chemically bound, their distance may be considered fixed, so that $\|P_j\|$ is constant, and the only dependence is on the orientation of the vector P_j . If the atoms belongs to the same domain as the metal, the tensor χ and the vector P_j belong to the same rigid structure. The RDC of the metal domain can be used to fit the numerical values of the χ tensor using (1).

The PCS measurements [18] are given by

$$(2) \quad \delta_{pcs,j} = \frac{c_{pcs}}{\|P_j\|^5} P_j^t \chi P_j,$$

a formula very similar to (1), with a different constant. In this case P_j is however the vector connecting the metal and selected atoms j . When the atom belongs to the metal domain there is no difference between RDC and PCS. In fact the atom does not move with respect to the metal ion, so P_j is fixed. Indeed RDC and PCS of atoms of the metal domain can be coupled to obtain a better fit for the paramagnetic tensor χ (and possibly for the location of the paramagnetic metal ion), see for instance [22]. From now on we suppose that this is the case, so that the paramagnetic tensor is known and only RDC and PCS for atoms belonging to the C terminal are considered. Since the C terminal moves with respect to the metal ion, the NMR measurements are averages of different states of the molecule, so that we may speak about *mean* PCS or RDC.

The RDC and PCS are in fact the average of the values obtained for different positions of the C terminal (also called conformers). Each conformer is identified by an Euler transformation $E \equiv (R, t)$, where R is a rotation and t a translation. Note however that the P_j of formulas (1) for RDC are difference of coordinates, so that the translations cancel, and we have

$$(3) \quad E(P_j) = R(P_j - t) \text{ for PCS, } \quad E(P_j) = RP_j \text{ for RDC.}$$

Note also that $\|E(P_j)\| = \|P_j\|$ does not depend on (R, t) in the case of RDC. Because of the linker we can always suppose $t_{min} \leq \|t\| \leq t_{max}$, so that the space of allowed Euler transformations is compact.

Let D be the space of probability distributions on this compact space. Each $d \in D$ is identified by the probability density $p(R, t) \geq 0$, such

that $\int_{R,t} p(R, t) dR dt = 1$. Then

$$(4) \quad \bar{\delta}_{rdc,j} = \frac{c_{rdc}}{\|P_j\|^5} \int_{R,t} p(R, t) (RP_j)^t \chi(RP_j) dR dt.$$

Since p does not depend on t for the RDC, we have $p = p(R)$ and $\int_R p(R) dR = 1$. Using the *mean paramagnetic tensor*

$$(5) \quad \bar{\chi} = \int_R p(R) R^t \chi R dR dt.$$

equation (4) becomes:

$$(6) \quad \bar{\delta}_{rdc,j} = \frac{c_{rdc}}{\|P_j\|^5} P_j^t \bar{\chi} P_j.$$

The same technique cannot be used for PCS because of the term $E(P_j) = R_i(P_j - t_i)$, so that

$$(7) \quad \bar{\delta}_{pcs,j} = c_{pcs} \int_{R,t} \frac{p(R, t)}{\|R(P_j - t)\|^5} (R(P_j - t))^t \chi(R(P_j - t)) dR dt.$$

Different metal ions M_k may be substituted in the same binding site belonging to the N terminal without influencing the fold of the protein [1]. We suppose that each set of measurements relative to metal M_k is obtained by averaging values relative to conformers, using the same probability distribution $d \in D$. Note the following proposition, see for instance [27].

Proposition 2.1. *Independent PCS and RDC measurements may be obtained from at most 5 different metal ions M_k .*

This is due to the fact there are at most 5 linearly independent paramagnetic tensors χ^k relative to metals k . A sixth tensor χ^6 may be written as a linear combination of the first five tensors. Hence PCS and RDC (and indeed mean PCS and RDC) relative to this sixth metal can be written as the same linear combination of the measurements relative to the first five metals, see (4) and (7).

In general we cannot determine d from the measurements of the moving terminal. The problem is in fact underdetermined. The target distribution d is a function of six variables, those defining the Euler transformation. If we only consider RDC, d is a function of the three variables identifying the rotation, be them unitary quaternions or Euler angles.

On the other hand we only have a finite number of measurements. Moreover, it is well known that the maximal number of independent RDC measurements from atoms of the C terminal is 25, see for instance [24]. Also, the information content of PCS is weak [3]. Hence, no

matter how many measurements are available, the distribution d cannot be recovered except in some trivial cases.

We now report some well known properties of RDC and PCS, see for instance [27, 2]. We first examine the case of the RDC measurements.

Proposition 2.2. *The maximal number of independent mean RDC measurements is 25.*

This result is the consequence of two different properties:

- (i) The maximal number of independent mean paramagnetic tensors is 5 (see Proposition 2.1).
- (ii) The maximal number of independent mean RDC for each metal is 5.

For a proof, see [21, Theorem 3.2].

More precisely, the independence of the RDC measurements is directly correlated to the independence of the mean paramagnetic tensors (12).

Proposition 2.3. *Let n be the number of independent mean paramagnetic tensors. Then the number of independent RDC measurements is $5n$.*

For the proof we refer to [23].

Proposition 2.1 holds also for PCS, thus the maximum number of independent metal ions is again 5. However, in principle each mean PCS is independent from the other, see formula (8). In practise if two atoms j_1 and j_2 are close, so are P_{j_1} and P_{j_2} . Hence the values of the averaged PCS from (7) are also close. Mathematically speaking we may observe that the values $\bar{\delta}_{pcs,j_1}$ and $\bar{\delta}_{pcs,j_2}$ are heavily correlated, so that the new information added by atom j_2 is very weak.

3. THE SIMPLEX ALGORITHM

3.1. Geometrical setting. Suppose we have $n \leq 5$ metal ions, and that χ^k are already given or determined via the RDC and PCS of the metal domain. Take any $d \in D$, we can calculate the mean RDC and PCS with the general formula

$$(8) \quad \bar{\delta}_j = c_j \int_{R,t} \frac{p(R,t)}{\|E(P_j)\|^5} E(P_j)^t \chi^{k_j} E(P_j) dR dt,$$

where $p(R,t)$ is the probability density of d at (R,t) , and $E(P)$ is defined by (3). The values P_j , c_j and $k_j \leq n$ depend on the choice of atoms and the type of measurement. The term $\|E(P_j)\|$ is constant for RDC. In the case of PCS, for physical reasons the distance between the metal ion and any other atom is anyway bounded away from 0. Hence

we may suppose that the measurements $|\bar{\delta}_j|$ are uniformly bounded. We can obtain a certain number of RDC and PCS measurements for each of the n metals, not necessarily referring to the same atoms. Let n_{rdc} be the total number of mean RDC, n_{pcs} be the total number of mean PCS, and let $n_{\text{meas}} = n_{\text{rdc}} + n_{\text{pcs}}$.

We can collect the measurements calculated from (12) in a vector, so that each $d \in D$ defines a point $\bar{\delta} \in \mathbb{R}^{n_{\text{meas}}}$. The key point of the geometrical approach is the projection from the space of finite distributions to the space of the measurements. Let Π be such a projection, we may also decompose Π into the RDC and PCS components:

$$(9) \quad \Pi(d) \equiv \begin{pmatrix} \Pi_{\text{rdc}}(d) \\ \Pi_{\text{pcs}}(d) \end{pmatrix} = \begin{pmatrix} \bar{\delta}_{\text{rdc},1} \\ \dots \\ \bar{\delta}_{\text{rdc},n_{\text{rdc}}} \\ \bar{\delta}_{\text{pcs},1} \\ \dots \\ \bar{\delta}_{\text{pcs},n_{\text{pcs}}} \end{pmatrix}.$$

Let

$$(10) \quad V = \{v \in \mathbb{R}^{n_{\text{meas}}} : v = \Pi(d), d \in D\}.$$

The set V is compact because the measurements are uniformly bounded. The set V is convex because if $v_i \in V$, $v_i = \Pi(d_i)$, $i = 1, 2$, then $\lambda d_1 + (1 - \lambda)d_2 \in D$, $\forall \lambda \in [0, 1]$, so that

$$(11) \quad \lambda v_1 + (1 - \lambda)v_2 = \Pi(\lambda d_1 + (1 - \lambda)d_2) \in V.$$

Each convex set is the convex hull of its extreme points (also called vertices), i.e. the points that cannot be reconstructed using a convex combination of different points of the set. Let $\Delta \subset D$ the set of finite probability distributions, and let $\hat{\Delta} \subset \Delta$ the set of probability distributions made by a single point. Because of the non-linearity, in general it is not true that each $\Pi(\hat{d})$ is a vertex of V , though we suspect this is the case in our setting. On the other hand, the set of vertices is a subset of $\Pi(\hat{d})$, since V is the set of convex combinations of these points. We do not need the property that each $\Pi(\hat{d})$ may be uniquely reconstructed, so we can nevertheless identify the set of vertices with $\Pi(\hat{d})$.

Proposition 3.1. *For each $d \in D$ there exists a $\tilde{d} \in \Delta$ such that $\Pi(\tilde{d}) = \Pi(d)$.*

Proof. By Carathéodory's theorem, each $\Pi(d) \in V$ may be reconstructed with a convex combination of at most $n_{\text{meas}} + 1$ vertices of

V. Let $\Pi(d) = \sum_{i=0}^{n_{\text{meas}}} p_i \Pi(\hat{d}_i)$, with $\hat{d}_i \in \hat{\Delta}$. Then $\tilde{d} = \sum_{i=0}^{n_{\text{meas}}} p_i \hat{d}_i \in \Delta$ is the required distribution. \square

Remark: Proposition 3.1 entitles us to work with finite distributions of probability without loss of generality. If $d \equiv (p_i, R_i, t_i) \in \Delta$, formula (8) may be rewritten as

$$(12) \quad \bar{\delta}_j = c_j \sum_i \frac{1}{\|E(P_j)\|^5} p_i E(P_j)^t \chi^{k_j} E(P_j).$$

Proposition 3.2. *There exists a $d \in \Delta$ such that $\Pi(d) = 0$.*

Proof. The proposition is proven in [14] for the case of RDC, and a constructive example with a finite distribution is given in [31]. Fix the origin of the Cartesian system in the binding site of the metal. Let $\tilde{d} \in D$ such that the translation t is fixed and the rotational part coincides with the Haar measure $H(R)$, see for instance [39]. Then $\Pi_{\text{rdc}}(\tilde{d}) = 0$. With these choices we also have $\Pi_{\text{pcs}}(\tilde{d}) = 0$. Fix a j relative to a PCS in formula (12). Let $\tilde{P}_j = P_j - t$, then $\|E(P_j)\| = \|R\tilde{P}_j\| = \|\tilde{P}_j\|$ for every rotation R because the metal is in the origin. Hence

$$(13) \quad \begin{aligned} \bar{\delta}_j &= c_j \frac{1}{\|\tilde{P}_j\|^5} \int_R (R\tilde{P}_j)^t \chi^{k_j} (R\tilde{P}_j) H(R) dR \\ &= c_j \frac{1}{\|\tilde{P}_j\|^5} \tilde{P}_j^t \left(\int_R R^t \chi^{k_j} R H(R) dR \right) \tilde{P}_j = 0. \end{aligned}$$

This is due to the fact that the integrand in parenthesis is the mean paramagnetic tensor, and its integral is 0 for the Haar measure [14]. The existence of a $d \in \Delta$ is then guaranteed by Proposition 3.1. \square

The dimension $N \leq n_{\text{meas}}$ of the set V is a key point which can be determined from the data. Using the results of the previous section, if we suppose that we have at least 5 independent RDC measurements for each of the n metal ions, then $N = 5n + n_{\text{pcs}}$. However, since the PCS are only marginally linearly independent, it is to be expected that there are directions where the set V is very thin, so that the effective determination of N should involve also some numerical considerations. In the supplementary information we analyse in detail the linear independence of the PCS versus the RDC measurements, and the consequences on the expected results.

3.2. Definition of the MAP. Let \bar{d} the true unknown distribution of probability. Then, given any Euler transformation (R, t) we define

$$(14) \quad p_{\max}(R, t) = \max_{d \in \Delta} \{p : (p, R, t) \in d \text{ and } \Pi(d) = \Pi(\bar{d})\}.$$

In other words given any conformer, identified by the Euler transformation (R, t) , we look for the maximal coefficient p that we can apply to this conformer in a convex combination such that the projection in V is the same as that of \bar{d} . Suppose $\Pi(\bar{d})$ belongs to the interior of V . Let $\Pi(\hat{d}) = \Pi(1, R, t)$ be the vertex corresponding to the position (R, t) . Consider the line passing through $\Pi(\hat{d})$ and $\Pi(\bar{d})$, the segment in between the two points belongs to V because of the convexity. Moreover, since $\Pi(\bar{d})$ is internal, there exists a point $\Pi(q) \in V$ on the continuation of the segment on the side of $\Pi(\bar{d})$. Then $\Pi(\bar{d})$ is the convex combination of $\Pi(q)$ and $\Pi(\hat{d})$, i.e. there exists a $p \in (0, 1)$ such that

$$(15) \quad \Pi(\bar{d}) = p\Pi(\hat{d}) + (1 - p)\Pi(q).$$

By definition we have $p_{\max}(R, t) \geq p$. The value p can be explicitly determined using the distances (i.e. the L^2 norms) in \mathbb{R}^N , in fact

$$(16) \quad \Pi(\bar{d}) = \frac{\|\Pi(\bar{d}) - \Pi(q)\|}{\|\Pi(\hat{d}) - \Pi(q)\|} \Pi(\hat{d}) + \frac{\|\Pi(\bar{d}) - \Pi(\hat{d})\|}{\|\Pi(\hat{d}) - \Pi(q)\|} \Pi(q).$$

The maximal p which verifies (15) is then obtained from the q with projection in V having the maximal distance from $\Pi(\bar{d})$. Because of the convexity, $\Pi(q)$ is the point on the boundary of V on the continuation of the segment connecting $\Pi(\hat{d})$ and $\Pi(\bar{d})$. Unfortunately, except in some trivial cases, there is no analytical procedure for determining if a point $\Pi(q)$ belongs to the boundary of V , so that we have to use an iterative procedure.

3.3. The simplex algorithm. Let $N \leq 5n + n_{\text{pcs}}$ be the dimension of V . By Carathéodory's theorem there are $N + 1$ vertices of the convex V such that

$$(17) \quad \Pi(\bar{d}) = \sum_{j=0}^N p_j \Pi(\hat{d}_j^0),$$

with $p_i \geq 0$ and $\sum_i p_i = 1$. Note again that we cannot suppose that $\bar{d} = \sum_i p_i \hat{d}_i^0$ because in general the solution is not unique, we can only recover the projection. Let $S_0 \subset V$ be the simplex formed by the convex combinations of the vertices $\Pi(\hat{d}_i^0)$. We may suppose S_0 is a simplex in \mathbb{R}^N , i.e. the vectors $\Pi(\hat{d}_i^0) - \Pi(\hat{d}_0^0)$ are linearly independent in \mathbb{R}^N . Since the set $\Pi(\hat{d})$ is connected we may choose S_0 so that $\Pi(\bar{d})$ is internal to S_0 , i.e. $p_i > 0 \forall i$ [21].

Now take any position (R, t) and let $\hat{d} = (1, R, t)$, look at Fig. 1 for reference. Take the line r through $\Pi(\hat{d})$ and $\Pi(\bar{d})$. Since $\Pi(\bar{d})$ is internal to S_0 there is a point $P_0 \in \partial S_0$ on r on the side opposite to

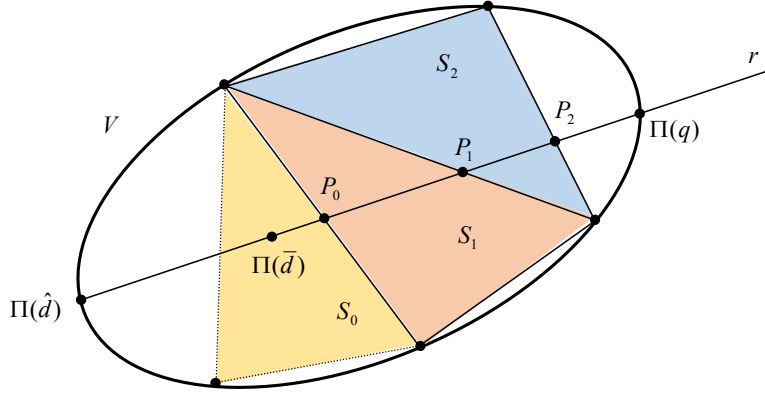


FIGURE 1. The simplex algorithm.

$\Pi(\hat{d})$ with respect to $\Pi(\bar{d})$. The point P_0 identifies a face $F_{j_0} \subset S_0$ such that $P_0 \in F_{j_0}$. The face F_{j_0} is identified by removing the vertex j_0 from the set of vertices of S_0 .

The point P_0 is either on the boundary or internal to V . In the first case we are finished because we have found the point needed by the definition of p_{\max} . In the second case, consider the hyperplane H_{j_0} containing the face F_{j_0} . The hyperplane H_{j_0} cannot be a support hyperplane since it contains an internal point, thus there will be at least a vertex $\hat{d}_{j_0}^1$ on the half-plane defined by H_{j_0} and not containing $\Pi(\bar{d})$. Define S_1 to be the simplex with $\hat{d}_{j_0}^0$ replaced by $\hat{d}_{j_0}^1$.

We can iterate the algorithm, each time finding the two intersections of r with the simplex S_k . The intersection point P_k on r farther from $\Pi(\bar{d})$ determines a face F_{j_k} of the simplex S_k . If this intersection point is internal we can replace the vertex j_k of S_k not belonging to F_{j_k} with a new one, lying in the half-space determined by the hyperplane H_{j_k} and not containing $\Pi(\bar{d})$.

Thus we determine a monotonic sequence of points $P_k \in r$ converging to a point P . The point P cannot be internal to V , otherwise the algorithm would have found a new replacement vertex. Then $P \in \partial V$ is the point needed by the definition of p_{\max} .

3.4. Determination of the projection matrix. In principle the algorithm may be carried out in the ambient space $\mathbb{R}^{n_{\text{meas}}}$ without any modifications. However, the dimension N of V is in general strictly smaller than the number of measurements n_{meas} . The simplex algorithm works in the linear subspace spanned by V , which has dimension N . Using the n_{meas} ambient coordinates in this linear subspace is

definitely a bad idea, because any numerical approximation in the calculations is likely to bring the points out of the linear subspace. Thus the first step is to determine the dimension N of V , and the projection operator from $\mathbb{R}^{n_{\text{meas}}}$ into \mathbb{R}^N .

The dimension N is the maximal number of linearly independent vectors of the form $\Pi(d_i) - \Pi(d_0)$, where d_0 is a fixed point in Δ , and $d_i \in \Delta$. Because of Proposition 3.2, we may take d_0 such that $\Pi(d_0) = 0$. Since each point in V is a convex combination of vertices, N is then the maximal number of linearly independent vectors $\Pi(\hat{d}_i)$, where $\hat{d}_i \in \hat{\Delta}$.

The Singular Value Decomposition (SVD, see for instance [26]) may be used to determine N , as already done in [32] in a different context. Take points $\Pi(\hat{d}_i) \in V$, $i = 1, \dots, M$, with $M \gg n_{\text{meas}}$, and form the matrix

$$(18) \quad A = (\Pi(\hat{d}_1), \dots, \Pi(\hat{d}_M)),$$

which has dimension $n_{\text{meas}} \times M$. The SVD is based on the singular values of A , which are the square roots of the eigenvalues of the symmetric and positive semi-definite matrix $A^t A$.

The SVD decomposes the matrix A in the form

$$(19) \quad A = U \Lambda W,$$

where Λ is a $n_{\text{meas}} \times n_{\text{meas}}$ diagonal matrix containing the singular values of A in decreasing order, W is an orthogonal $n_{\text{meas}} \times n_{\text{meas}}$ matrix, and U is a $M \times n_{\text{meas}}$, column-orthogonal matrix, i.e. $\sum_k u_{ki} u_{kj} = \delta_{ij}$. Since the rank of A is by definition N , the matrix A has exactly N non-zero singular values.

The matrix W can be used as a projection matrix. In fact, for any $v \in V$ we have that Wv is the decomposition of v with respect to the base of eigenvectors of $A^t A$. Moreover, since the matrix Λ has only N non-zero eigenvalues, we are only interested in the matrix $W_N \subset W$, including only the first N rows of W .

While the matrix $W_N^t W_N$ is not the identity, it works as such on the points of V . In fact the points of the linear space spanned by V can be uniquely identified either by a subset of the n_{meas} coordinates satisfying $n_{\text{meas}} - N$ linear conditions, or by the N intrinsic coordinates obtained by applying W_N .

The following diagram pictures the situation:

$$\begin{array}{ccccccc} \Delta & & \mathbb{R}^{n_{\text{meas}}} & & \mathbb{R}^N & & \mathbb{R}^{n_{\text{meas}}} \\ & \Pi & & W_N & & W_N^t & \\ d & \rightarrow & \Pi(d) & \rightarrow & W_N(\Pi(d)) & \rightarrow & \Pi(d) \end{array}$$

4. IMPLEMENTATION AND SELF-CONSISTENCY TESTS

In this section we describe the implementation of the simplex algorithm and report the results of simulations run with exact measurements. It is a "proof of concept" that the method works, and a necessary step to understand the interactions of the measurements before considering noisy data.

4.1. Experimental setting. As a model for calmodulin we use the pdb fold as determined by [8], shown in Figure 2. In physiological

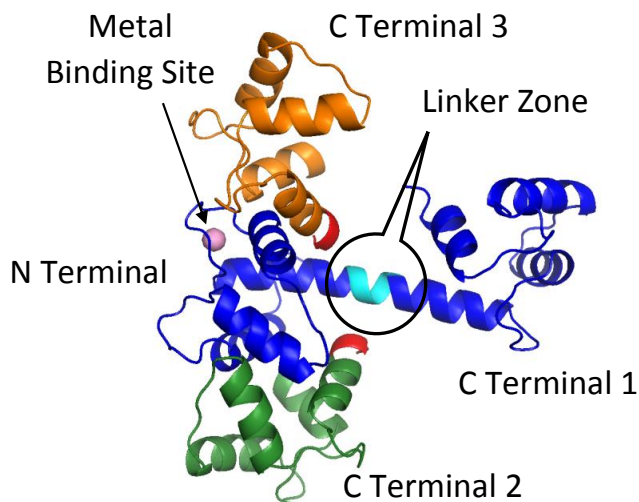


FIGURE 2. The fold of calmodulin and its conformational freedom. For the figure in color: *pink*: metal position; *blue*: N terminal and position C_1 of the C terminal, linker (*cyan*) shown only for this position; *green*: position C_2 ; *gold*: position C_3 . Residuals 82–83 of C terminal shown in *red* for C_2 and C_3 in order to highlight the connection points of the linker(not shown) to the C terminal.

conditions the long α -helix breaks in the zone between residuals 77 and 81 [6, 5], resulting in some conformational freedom of the C terminal. Any position such that the linker length is between 6\AA and 12\AA is considered to be attainable. Any position such that there exists two C_α atoms with a distance smaller than 3\AA is considered to be a physical violation. A conformer satisfying both conditions is an allowable state

for the C terminal. The positions of the C terminal shown are marked C_1 through C_3 , with C_1 being the position where the linker remains folded into the α -helix. In position C_2 the C terminal is close to the N terminal, but far from the metal. In position C_3 the C terminal is close both to the N terminal and to the metal. We have highlighted in red the first two residuals of the C terminal for positions C_2 and C_3 to show the connection point to the linker.

The measurements are generated using a continuous probability distribution $d_l \in D$ centered in conformers $C_l \equiv (R_l, t_l)$, $l = 1, 2, 3$. The measurements are calculated using formula (12) as the arithmetic average of a large number M of allowable states drawn according to the distribution d_l . More precisely, given two positive numbers σ_R and σ_t we draw conformers in the following way. The translation is drawn according to a Gaussian distribution with average t_k and standard deviation of the module σ_t . The rotation is drawn according to a von Mises-Fisher distribution (see for instance [39]) with average R_k and standard deviation σ_R of the rotational distance, calculated using quaternions. We only retain allowable states. The number M is large enough to stabilize the measurements, thus simulating a continuous probability distribution. Loosely speaking the C terminal moves around the center position C_k in such a way that the average deviation from the central position is σ_t Å for the translation, and σ_R degrees for the rotation. In this paper we use the numerical values $\sigma_t = 3\text{Å}$ and $\sigma_R = 20^\circ$.

We generated mean measurements with respect to three different paramagnetic tensors, corresponding to Tb, Tm and Dy lanthanide ions substituted for Ca in the second binding site of the N-terminal. We simulated a total of 112 mean RDC using N-H dipoles from residuals of the C terminals, and a total of 132 mean PCS using HN atoms from the C terminal.

In principle the distribution is symmetric around the center. However the constraint on the physical violations may introduce asymmetries in the distribution. This happens in cases C_2 and C_3 , when the center position of the C terminal is close to the N terminal. As a consequence there is a small shift in the most probable position of the distribution. In the supplementary information we discuss in detail this issue.

4.2. Observability of a central tendency. Suppose there is a central tendency in the data. The MAP estimate is able to detect whether this tendency exists. To show this fact, we considered a probability distribution d_0 where all the allowable states are equally probable in

the correct metric. If there were no constraints on the conformers the simulated measurements would be 0 by Property 3.2. The large conformational freedom is however detectable from the RDC measurements. The standard deviation of the simulated RDC is in fact 0.36 for d_0 , while is larger than 5 for the d_l cases, $l \geq 1$. The situation is different for PCS. In this case, small values are obtained both when there is a large conformational freedom and when the C terminal is far from the binding site of the metal. The standard deviation of the simulated PCS is 0.10 for d_0 , 0.14 for d_1 , 0.08 for d_2 and finally 0.94 for d_3 , the case where the C terminal is closer to the metal binding site.

The MAP estimate detects this difference in the data. In the d_0 case using RDC we found $0.31 \leq p_{\max}(R) \leq 0.34$ for all the orientations of the C terminal. Typical values for the other d_i cases are $0.1 \leq p_{\max}(R) \leq 0.70$, thus having a much larger span. The MAP estimate is then in principle able to detect an asymmetry in the data due to a restricted conformational freedom.

4.3. Determination of the central tendency. We used the following steps to determine a central tendency of the measurement.

1. The orientation R_0 of the conformer with the largest p_{\max} is determined using RDC alone.
2. The translation t_0 of the conformer with the largest p_{\max} is determined using PCS alone. The rotation is kept fixed at $R = R_0$.

In the following figures we present the results of the tests.

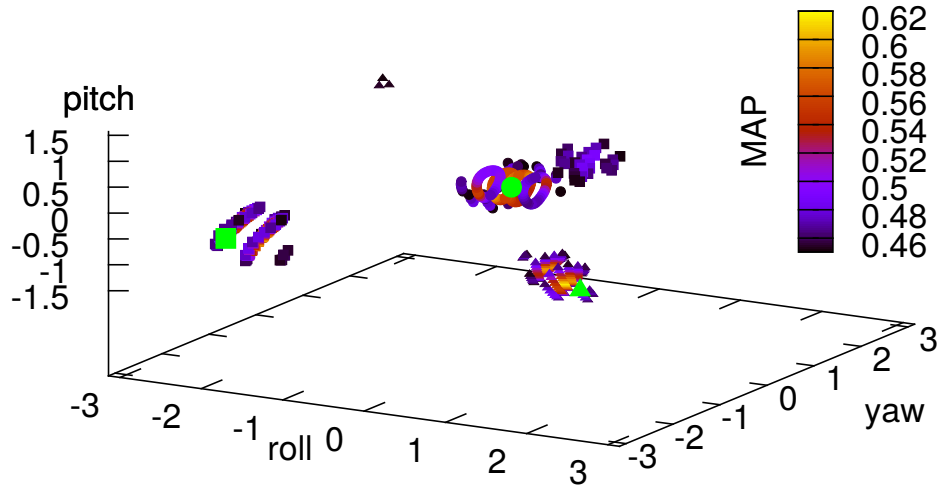


FIGURE 3. Tests with RDC alone. Cases C_1 (circles), C_2 (squares), C_3 (triangles).

In Figure 3 we show the results of step 1 of the algorithm in cases C_1 (circles), C_2 (squares), C_3 (triangles). The points represent the orientations of the sample with MAP larger than a certain threshold. The larger green dots, here and in the subsequent figures, mark the positions of the centers. In cases C_2 and C_3 there are two different zones with large MAP, due to the so-called phenomenon of *ghost cones* [21], which derives from the symmetries of the RDC formula.

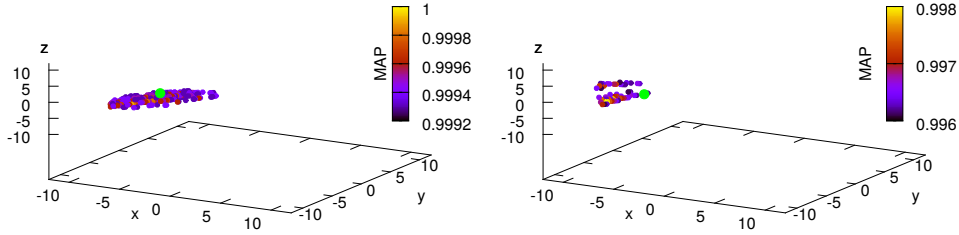


FIGURE 4. Tests with PCS alone. Cases C_1 (left panel) and C_3 (right panel).

In Figure 4 we show the results of step 2 of the algorithm in cases C_1 and C_3 , case C_2 being very similar to case C_1 . Note that the MAP of a large number of translations is close to 1, as a consequence of the poor resolving power of the PCS. A slightly better reconstruction is obtained for C_3 . In this case the center position is close to the metal, so that there are some PCS values which are rather large, see formula (7). To obtain these values the C terminal must remain close to the C_3 position for a not negligible fraction of time. As a consequence, the p_{\max} values for positions far from the metal is slightly reduced.

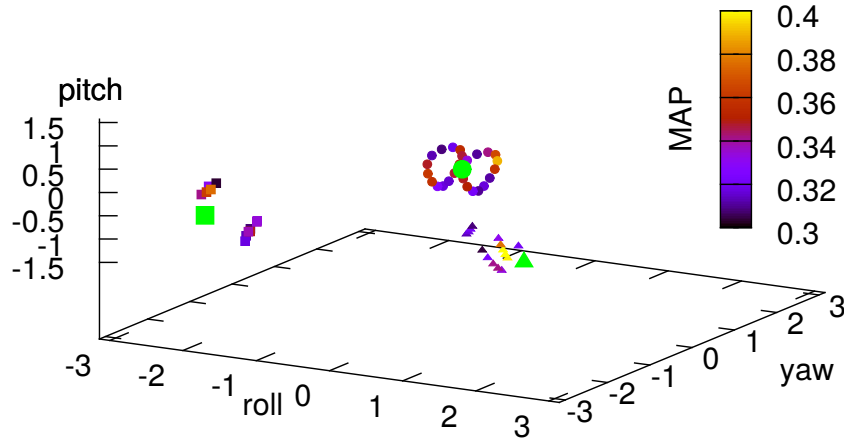


FIGURE 5. Tests with RDC and PCS. Cases C_1 (circles), C_2 (squares), C_3 (triangles).

In Figure 5 we show the joint PCS+RDC case for determining an orientation. While the PCS do not resolve well the translation, they are useful to eliminate the ghost cones, thus determining the correct region of space for the orientations. The RDC and PCS formulas have the same type of symmetries, however the $E(P_j)$ vectors from (3) are different, so that in general the symmetries do not coincide.

5. TESTS WITH EXPERIMENTAL ERROR

We added an uncorrelated Gaussian error to the mean measurements to take into account the experimental error. The error level was kept to $\pm 1\text{ppm} \pm 10\%$ for PCS and $\pm 1\text{Hz} \pm 10\%$ for RDC. We applied the algorithm of Subsection 4.3 and report the results in the following figures.

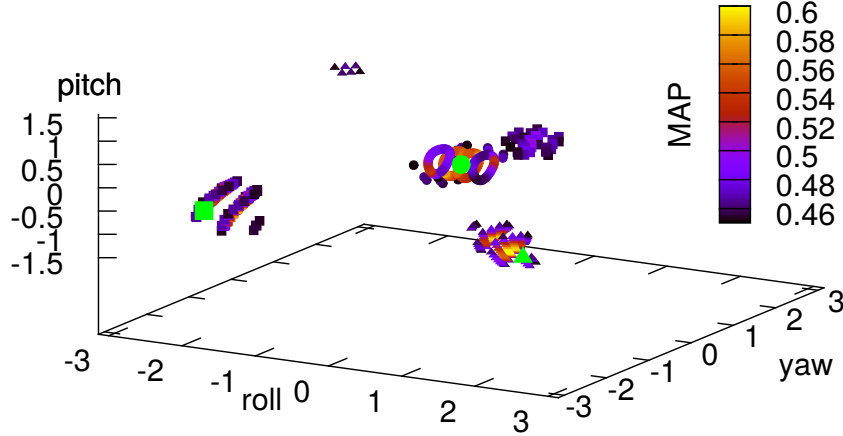


FIGURE 6. Tests with RDC alone. Cases C_1 (circles), C_2 (squares), C_3 (triangles).

In Figure 6 we show the results of step 1 of the algorithm. Note that there are very few changes with respect to Figure 3. This is due to the fact that the mean RDC have only 15 degrees of freedom, while we have 112 measurements. The SVD algorithm implicitly fits the 15 degrees of freedom of the mean RDC. Since the error is assumed to be Gaussian, a large number of measurements reduces the standard error of the fitted quantities. Hence the information of the RDC is well recovered even when the experimental error is considered. As explained in the previous section, coupling PCS and RDC helps removing ghost cones.

In Figure 7 we show the results of step 2 of the algorithm, in case C_1 (left panel) and C_3 (right panel). Here the introduction of the experimental error worsen the results. This should not be a surprise,

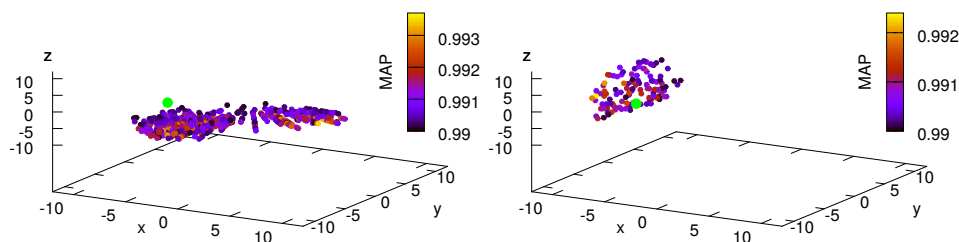


FIGURE 7. Tests with PCS alone. Cases C_1 (left panel) and C_3 (right panel).

since most of the mean PCS are under the 1Hz threshold, so that their numerical value is destroyed by the error level which is larger. As already explained, the case C_3 is better because the C terminal is close to the metal.

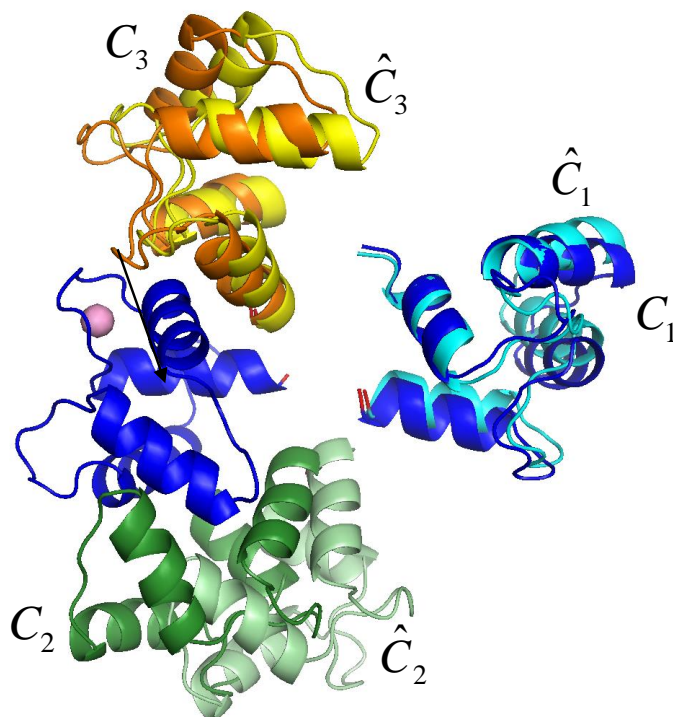


FIGURE 8. Tests with RDC and PCS. Metal (*pink*), N terminal (*blue*). Final reconstruction of the C_1 , C_2 and C_3 cases. Dark colors show the positions with the largest probability, light colors show the reconstruction.

Figure 8 shows the final reconstruction. Here the dark colored conformers show the positions with the largest probability, while the lightly colored conformers show the reconstructed positions \hat{C}_k .

It should be noted that the algorithm does not include a step where PCS and RDC are analysed together, if not in order to remove the ghost cones. We did attempt this step, using a local maximization technique. The results of this step are on average only marginally better than the results of the algorithm. If great care is not applied in the optimization, the error on the translation may even increase. Even in the joined RDC+PCS case, the translation is determined only by the PCS values, since there is no dependence on the RDC. However, the dependence on the translation is very weak, as it can be seen from figure 7, so that the variations in the MAP are largely due to the orientation of the conformer. More information on this additional attempt may be found in the supplementary material.

6. CONCLUSIONS

In this paper we demonstrated the ability of PCS and RDC measurements to recover a central tendency of an unknown probability distribution representing the conformational freedom of a protein made by two rigid domains connected by a flexible linker.

We made use of the MAP algorithm, extended to include PCS (and indeed any other class of measurements). The MAP algorithm determines the largest probability of a conformer in a distribution satisfying the measurements. Taken globally the MAP function is not a probability distribution but a sharp bound from above. For each position there is however an explicitly determined finite probability distribution with the MAP value as a weight for that conformer.

The RDC measurements are well able to determine any central tendency in the orientation of the conformer. Adding the PCS helps removing symmetric orientations, since the symmetries of the PCS are different from those of the RDC.

The identification of the translation is more difficult. The RDC does not depend on the translation, so we can only use PCS. However the information content of the PCS is very weak, and is further destroyed by the experimental noise. With exact data we can only approximatively determine the central tendency of the translation, see Figure 4. The situation worsen when the experimental error is added, as shown by Figure 7. Values of MAP larger than 0.9 are obtained for a large fraction of the sample, and especially for positions relatively far from the metal. In other words, the conformer can sample any positions far from the metal for the 90% of the time, resulting in very small partial values of the PCS. Adjusting the remaining 10% of the distribution is

enough to obtain the correct values of the PCS. As a consequence, the determination of the translation is less accurate.

We again stress that this is not due to the method employed. The MAP algorithm points out the extremal cases which should anyway be considered by any method trying to determine a solution. Of course there might be reasons to exclude some probability distributions, for instance using some threshold on the number of conformers or on the spread of the distribution. These are additional hypotheses which reduce the set of compatible solutions. However, since the problem is underdetermined and the real solution is not known, the reasons for removing these particular solutions should be soundly justified.

Other methods might not be able to detect the MAP solutions, for instance if a predetermined sample is used for the conformers. Even if the sample is large, there might be a correlation between the rotational and translational part of the Euler transformations of the sample. In this case the identification of the orientation might bias the translation towards the correct value.

Further developments will include the case when there is more than a single central tendency in the data. Giving the results of this study, we foresee that additional information might be extracted only for the rotational part of the distribution. A possible way of overcoming this difficulty might be the inclusion of different measurements such as SAXS (Small angle X-Rays Scattering) [33]. The NMR-SAXS integration has already been studied using the MO (Maximum Occurrence) method in [9]. Since the SAXS measurements depend on the global shape of the molecule this information might help in better determining the translational part of the Euler transformation.

Supplementary Information: In the supporting material some additional results are presented. While not strictly necessary for the purpose of the paper, these results increase the insight on the behaviour of the PCS and RDC class of measurements.

Acknowledgment: As usual we wish to acknowledge the fruitful and long-established collaboration with the Center for Magnetic Resonance of the University of Florence. Discussions with applied scientists is always fruitful and helps gearing mathematics towards realistic problems.

REFERENCES

- [1] Allegrozzi M, Bertini I, Janik M B L, Lee Y-M, Liu G and Luchinat C 2000 Lanthanide induced pseudocontact shifts for solution structure refinements of macromolecules in shells up to 40 Å from the metal ion *J. Amer. Chem. Soc.* **122** 4154–4161

- [2] Al-Hashimi H M, Valafar H, Terrel M, Zartler E R, Eidsness M K and Prestegard J H 2000 Variation of molecular alignment as a means of resolving orientational ambiguities in protein structures from dipolar couplings *J. Magn. Reson.* **143** 402–406
- [3] Andralojc W, Berlin K, Fushman D, Luchinat C, Parigi P, Ravera E, Sgheri L 2105 Information content of long-range NMR data for the characterization of conformational heterogeneity *J. Biomol. NMR* **62** 353–371
- [4] Berlin K, Castaneda C A, Schneidman-Duhovny D, Sali A, Nava-Tudela A and Fushman D 2010 Recovering a Representative Conformational Ensemble from Underdetermined Macromolecular Structural Data *J. Amer. Chem. Soc.* **122** 4154–4161
- [5] Baber J, Szabo A and Tjandra N 2001 Analysis of Slow Interdomain Motion of Macromolecules Using NMR Relaxation Data, *J. Amer. Chem. Soc.* **123** 3953–3959
- [6] Barbato G, Ikura M, Kay L E, Pastor R W and Bax A 1992 Backbone dynamics of calmodulin studied by ^{15}N relaxation using inverse detected two-dimensional NMR spectroscopy: the central helix is flexible. *Biochemistry* **31** 5269–5278
- [7] Bertini I, Del Bianco C, Gelis I, Katsaros N, Luchinat C, Parigi G, Peana M, Provenzani A and Zoroddu M A 2004 Experimentally exploring the conformational space sampled by domain reorientation in calmodulin *Proc. Natl. Acad. Sci. USA* **101** 6841–6
- [8] Bertini I, Gelis I, Katsaros N, Luchinat C, Provenzani A 2003 Tuning the affinity for lanthanides of calcium binding proteins *Biochemistry* **42** 8011–8021
- [9] Bertini I, Giachetti A, Luchinat C, Parigi G, Petoukhov MV, Pierattelli R, Ravera E and Svergun DI 2010 Conformational Space of Flexible Biological Macromolecules from Average Data *J. Am. Chem. Soc.* **132** 13553–13558
- [10] Bertini I, Gupta K J, Luchinat C, Parigi G, Peana M, Sgheri L and Yuan J 2007 Paramagnetism-Based NMR Restraints Provide Maximum Allowed Probabilities for the Different Conformations of Partially Independent Protein Domains *J. Am. Chem. Soc.* **129** 12786–12794
- [11] Cavalli A, Camilloni C, and Vendruscolo M 2013 Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle *J. Chem. Phys.* **138** 094112
- [12] Evans M, Hastings N and Peacock B 2000 *Statistical distributions*, 3rd ed. (New York: Wiley)
- [13] Frauenfelder H, Sligar S and Wolynes P 1991 The energy landscapes and motions of proteins *Science* **254** 1598–1603
- [14] Gardner R, Longinetti M, and Sgheri L 2005 Reconstruction of orientations of a moving protein domain from paramagnetic data *Inverse Problems* **21**, 879–898
- [15] Huang Y J and Montelione G T 2005 Structural biology: proteins flex to function *Nature* **438** 36–37
- [16] Ikegami T, Verdier L, Sakhaei P, Grimme S, Pescatore B, Saxena K, Vogtherr M, Fiebig K M and Griesinger C 2004 Novel techniques for weak alignment of proteins in solution using a chemical tags coordinating lanthanide ions, *J. Biomol. NMR* **29** 339–349.

- [17] Jaynes E 1979 Where do we stand on maximum entropy? In: *Levine R, Tribus M, editors. The Maximum Entropy Formalism*. MIT Press (Cambridge MA) 1–104
- [18] Kurland J R and McGarvey B R 1970 Isotropic NMR shifts in transition metal complexes: the calculation of the Fermi contact and pseudocontact terms, *J. Magn. Reson.* **2** 286–301
- [19] Kullback S and Leibler R A 1951, On Information and Sufficiency *Ann. Math. Statist.* **22** 79–86
- [20] Lindorff-Larsen K, Best R B, De Pisto M A, Dobson C M and Vendruscolo M 2005 Simultaneous determination of protein structure and dynamics *Nature* **433** 128–132
- [21] Longinetti M, Luchinat C, Parigi G and Sgheri L 2006 Efficient determination of the most favoured orientations of protein domains from paramagnetic NMR data *Inverse Problems* **22**, 1485–1502
- [22] Longinetti M, Parigi G and Sgheri L 2002 Uniqueness and degeneracy in the localization of rigid structural elements in paramagnetic proteins *J. Phys. A* **35** 8153–69
- [23] Longinetti M, Sgheri L and Sottile F 2010 Convex Hulls of Orbits and Orientations of a Moving Protein Domain *Discrete Comput. Geom.* **43** 54–68
- [24] Meiler J, Prompers J J, Peti W, Griesenger C and Brüshweiler R 2001 Model-free approach to the dynamic interpolation of residual dipolar coupling in globular proteins *J. Am. Chem. Soc.* **123** 6098–6107
- [25] Palmer A G, Massi F 2006 Characterization of the dynamics of biomacromolecules using rotating-frame spin relaxation NMR spectroscopy *Chem. Rev.* **106** 1700–1719
- [26] Press W H, Teukolsky S A, Vetterling W T and Flannery B P 1992 *Numerical recipes in C: the art of scientific computing (second edition)* (New York: Cambridge Univ. Press)
- [27] Ramirez B E and Bax A 1998 Modulation of the Alignment Tensor of Macromolecules Dissolved in a Dilute Liquid Crystalline Medium *J. Am. Chem. Soc.* **120** 9106–9107
- [28] Rodruéz-Castañeda F, Maestre-Martínez M, Coudevylle N, Dimova K, Junge H, Lipstein N, Lee D, Becker S, Brose N, Jahn O, Carlomagno T and Griesinger C 2010 Modular architecture of Munc13/calmodulin complexes: dual regulation by Ca^{2+} and possible function in short-term synaptic plasticity *EMBO Journal* **29** 680–691
- [29] Ravera E, Sgheri L, Parigi G, Luchinat C 2015 A critical assessment of methods to recover information from averaged data, *Phys. Chem. Chem. Phys.* doi: 10.1039/C5CP04077A
- [30] Salmon L and Blackledge M 2015 Investigating protein conformational energy landscapes and atomic resolution dynamics from NMR dipolar couplings: a review, *Rep. Prog. Phys.* **78** 126601
- [31] Sgheri L 2010 Joining RDC data from flexible protein domains *Inverse Problems*, **26** 115021
- [32] Sgheri L 2010 Conformational freedom of proteins and the maximal probability of sets of orientations *Inverse Problems*, **26** 035003

- [33] Svergun D I, Barberato C, Koch M H J 1995 CRY SOL – a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates, *J. App. Cryst.* **28** 768–773
- [34] Selkoe D J 2003 Folding proteins in fatal ways *Nature* **426** 900–904
- [35] Schneider R and Weil W 1992 *Integralgeometrie* (Stuttgart: B G Teubner)
- [36] Tolman J R, Al-Hashimi H M, Kay H M, and Prestegard J H 2001 Dynamic Analysis of Residual Dipolar Coupling Data for Proteins *J. Am. Chem. Soc.* **123** 1416–1424
- [37] Tolman J R, Flanagan J M, Kennedy M A, and Prestegard J H 1995 Nuclear magnetic dipole interactions in field-oriented proteins: information for structure determination in solution *Proc. Natl. Acad. Sci. USA* **92** 9279–83
- [38] Van Kampen N G 1981 *Stochastic processes in physics and chemistry* (New York: North-Holland)
- [39] Weisstein E W 1998 *CRC Concise Encyclopedia of Mathematics* (Boca Raton: Chapman and Hall)

⁽¹⁾ISTITUTO PER LE APPLICAZIONI DEL CALCOLO (CNR), SEDE DI FIRENZE,
VIA MADONNA DEL PIANO, 10, 50019 SESTO FIORENTINO (FI), ITALY
E-mail address: *l.sgheri@iac.cnr.it